Document: T10/01-177r0 Date: 21 May 2001
To: T10 Committee Membership
From: Edward A. Gardner, Ophidian Designs
Subject: SRP Model for RDMA Communication Services

Attached is a "replacement" for subclause 4.1 of SRP revision 04. Upon re-writing this, it became clear to me that this material should be entirely informative. It defines or explains what it is to be an RDMA communication service. Other than terminology definitions, it places no constraints upon such a service. It places no constraints upon SRP. My expectation is that normative statements will appear elsewhere, such as in the SRP to SAM-2 mapping section.

My understanding is that purely informative material is supposed to appear in annexes. Therefore I have formatted this as an annex. The specific proposal is to remove subclause 4.1 of SRP revision 04 in its entirety and add this as Annex A. I welcome comments on whether there is a more appropriate way to present this material.

Annex A

(informative)

RDMA communication services

A.1 Overview

SRP is designed to operate using an RDMA communication service. RDMA communication services provide messages for control information and remote direct memory access (RDMA) operations for data transfers.

Figure A.1 shows an example system that uses an RDMA communication service. Communication is provided by RDMA channels, shown as solid lines in the figure. An RDMA channel provides communication between two consumers. A single pair of consumers may communicate using many RDMA channels if sufficient resources are available. In some environments it may be desirable to use multiple RDMA channels with specialized purposes. For example, a pair of consumers could use certain RDMA channels for messages (control information) and other RDMA channels for RDMA operations (data transfers).

The RDMA communication service in figure A.1 is shown as comprised of adapters and other unspecified components (e.g. wires, fabric switches, etc.). The actual components of an RDMA communication service depend on its implementation. Specific components such as adapters may or may not be present.

An RDMA communication service or a specific implementation of one may or may not provide the exact functions described here. A function described here might be mapped to a sequence of several functions provided by the RDMA communication service. An annex documenting the use of SRP on an RDMA communication service describes the mapping of functions described here to functions provided by that RDMA communication service.

A.2 RDMA Channels

An RDMA channel provides message communication and/or RDMA operations between a pair of consumers. An RDMA channel is a dynamic connection, established and disconnected upon request. Establishing an RDMA channel may require obtaining resources to support the channel, either within the channel's consumers or within

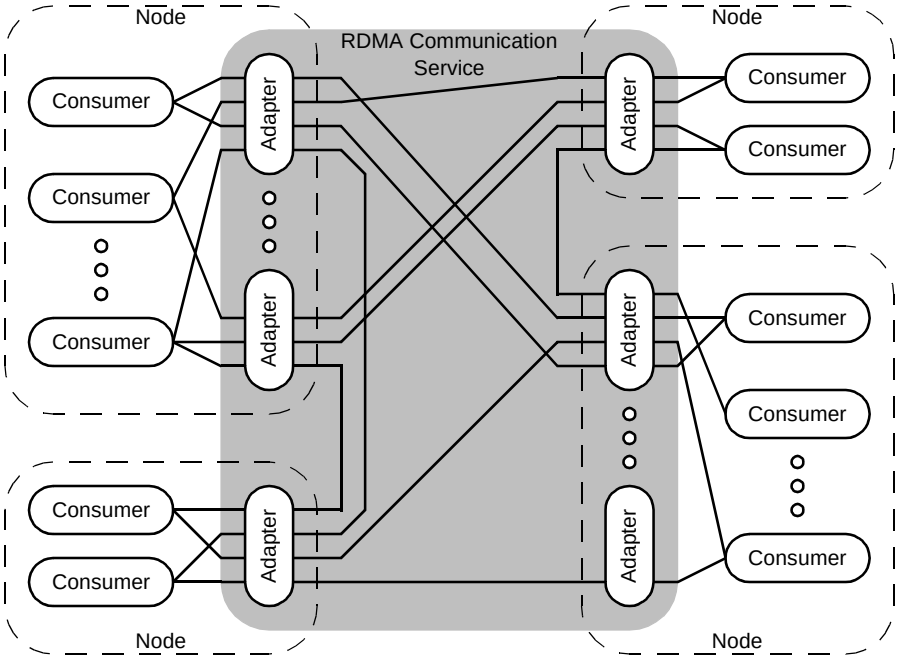

Figure A.1 - RDMA Communication Service Example

the RDMA communication service or both. Multiple channels may be established between the same pair of consumers if sufficient resources are available. The resources associated with a channel may be released after the channel is disconnected.

Figure A.2 shows the process by which a new RDMA channel is established. A client consumer requests that the RDMA communication service establish a new RDMA channel. The request is directed to a server and, if successful, resolved to a server consumer. The resulting RDMA channel provides communication between the client consumer and the server consumer.

A client consumer provides a SERVER ADDRESS to identify the server with which to establish an RDMA channel. The format and interpretation of a SERVER ADDRESS are specific to the RDMA communication service. A SERVER ADDRESS may specify an individual server consumer. A SERVER ADDRESS may also specify multiple server consumers. For example, a SERVER ADDRESS might identify an adapter as shown in figure A.1, specifying all consumers that implement a specific application protocol and are accessible through that adapter.

In figure A.2 the target of an RDMA channel establishment request, identified by a SERVER ADDRESS, is called a server agent. The server agent uses application protocol and/or server specific knowledge to determine whether an RDMA channel establishment request can be accepted and the server consumer to which it will be assigned.

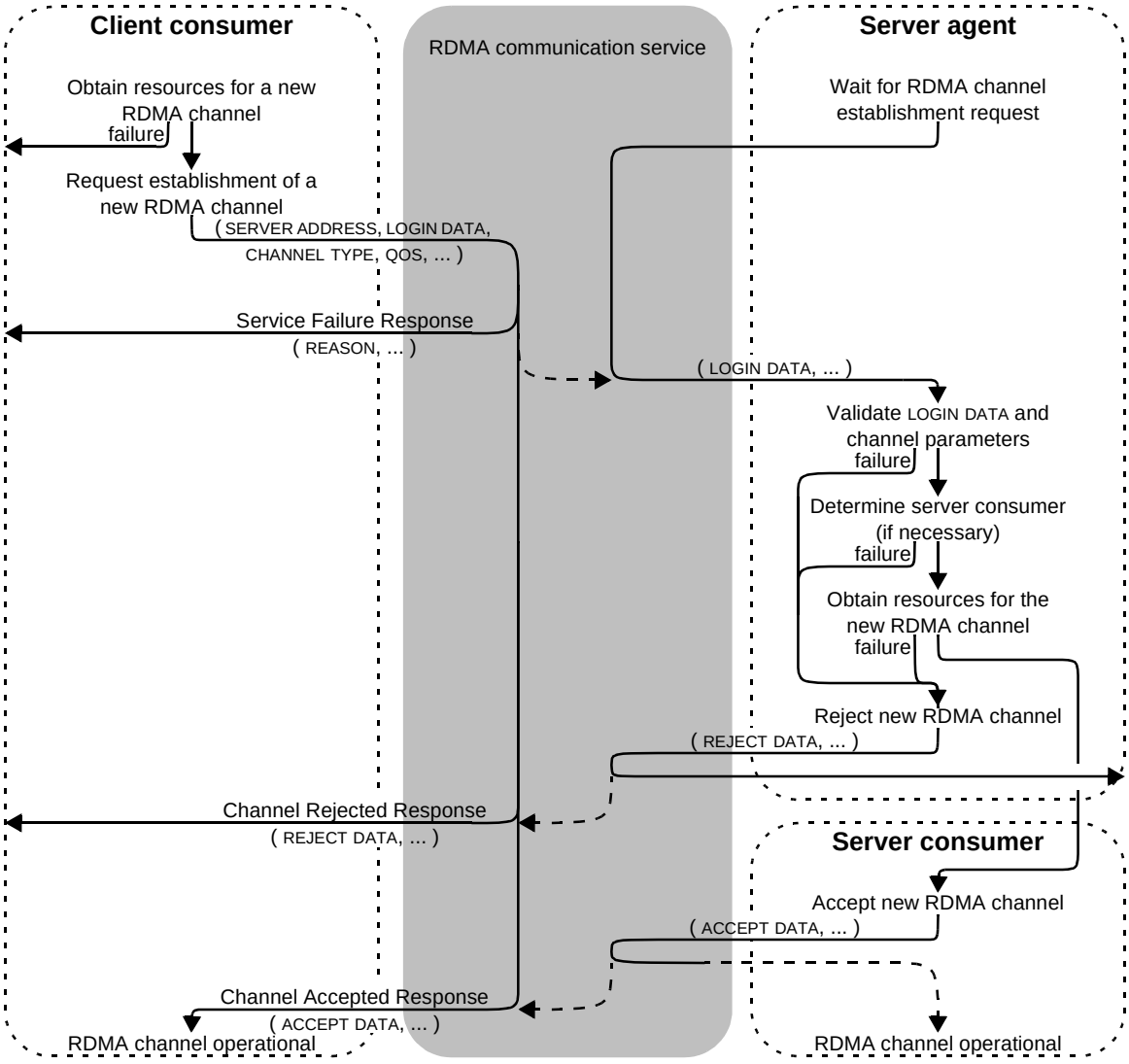

Figure A.2 - RDMA channel establishment

The actions shown as performed by a server agent in figure A.2 and the order in which they are performed are informative; the actions required of a server agent are specific to the RDMA communication service and/or server. A server agent may or may not be a distinct entity. Some or all of the actions of a server agent may be performed by a server consumer or by the RDMA communication service.

An RDMA communication service may require that the client consumer obtain certain resources before requesting that a new RDMA channel be established. After obtaining any such resources, the client consumer may request that the RDMA communication service establish a new RDMA channel. The request includes:

a. SERVER ADDRESS: identifies the server to which the channel should be established.

b. LOGIN DATA: application protocol data that may be used by the server agent or consumer.

c. CHANNEL TYPE: identifies the type of RDMA channel desired, if the RDMA communication service supports multiple types.

d. QOS: identifies the desired performance characteristics of the RDMA channel.

e. Any other parameters required by the RDMA communication service.

The format and interpretation of all parameters except LOGIN DATA are specific to the RDMA communication service.

The RDMA communication service returns one of three responses to the client consumer for a new RDMA channel establishment request.

The RDMA communication service may return a Service Failure Response. This indicates that the new RDMA channel could not be established for some reason internal to the RDMA communication service. A Service Failure Response may return a REASON to identify the cause of the failure as well as other RDMA communication service specific data.

The RDMA communication service may return a Channel Rejected Response. This indicates that the request was communicated to the server but rejected by the server agent or consumer. A Channel Rejected Response returns REJECT DATA, which is application protocol data provided by the server agent or consumer. REJECT DATA may include a reason for rejecting the request or other application protocol information. A Channel Rejected Response may also return RDMA communication service specific data.

The RDMA communication service may return a Channel Accepted Response. This indicates that the new RDMA channel has been successfully established and is operational. The client consumer may use the channel in accordance with the application protocol. A Channel Accepted Response returns ACCEPT DATA, which is application protocol data provided by the server agent or consumer. ACCEPT DATA may include application protocol parameters governing how the new RDMA channel should be used. A Channel Accepted Response may also return RDMA communication service specific data.

A server agent indicates its willingness to consider new RDMA channel establishment requests by registering with the RDMA communication service. In figure A.2 this is shown as a registration request (e.g. subroutine call) that returns control to the server agent when a new RDMA channel establishment request is received. That is informative; the way that a server agent registers with an RDMA communication service is specific to that service or the server.

New RDMA channel establishment requests that are acceptable to the RDMA communication service are passed to the server agent. The server agent is provided the LOGIN DATA from the client consumer's request as well as any RDMA communication service specific data that may be provided.

The server agent determines whether the new RDMA channel establishment request can be accepted and, if necessary, determines the server consumer that will be associated with the new RDMA channel. If the request cannot be accepted the server agent or consumer instructs the RDMA communication service to reject the new RDMA channel. The server agent or consumer provides REJECT DATA, which is application protocol data to be

passed to the client consumer, as well as any RDMA communication service or server specific data that may be required.

If the new RDMA channel establishment request can be accepted, the server agent or consumer instructs the RDMA communication service to accept the new RDMA channel. The server agent or consumer provides ACCEPT DATA, which is application protocol data to be passed to the client consumer, as well as any RDMA communication service or server specific data that may be required.

An RDMA channel is disconnected by a request from either of its consumers or from the RDMA communication service itself. The consumers will each be notified that the RDMA channel has been disconnected, allowing them to recover any resources associated with the RDMA channel. The delay until such a notification is given may vary depending upon the RDMA communication service, the node containing the consumer and the specific circumstances of the disconnection request.

A disconnect request causes an RDMA channel to become non-operational. Operations in progress on an RDMA channel at the time of a disconnect request and operations requested subsequent to a disconnect request may or may not complete.

A.3 Messages

An RDMA channel may provide message communication between its consumers. A message contains some number of payload bytes. A message is sent by one consumer associated with an RDMA channel, the sending consumer, to the other consumer associated with the RDMA channel, the receiving consumer. An RDMA communication service may provide normal and solicited message reception notification, which may be used to distinguish between more and less urgent messages.

A sending consumer requests that a message be sent by providing to an RDMA communication service:

a. The message's length.

b. Its payload contents.

c. The RDMA channel to use.

d. Whether to use normal or solicited message reception notification.

The RDMA communication service then attempts to deliver the message to the receiving consumer. If it succeeds, it notifies the receiving consumer that a message has been received, providing the message's length, payload contents and the channel on which it was received. The RDMA communication service may also provide an indication of whether the sending consumer specified normal or solicited message reception notification.

An RDMA communication service may require that receiving consumers provide message receive buffers to RDMA channels before messages are sent to them, and that the provided message receive buffers be large enough to hold any messages that arrive. Sending a message on an RDMA channel when no receive buffer has been provided, or when the provided receive buffer is too small for the message, may result in behavior that is specific to the RDMA communication service and/or the node(s) containing the consumer(s). Such behavior may include (but is not limited to):

a. Disconnecting the RDMA channel.

b. Discarding or truncating the message, with or without an error indication.

c. Delaying delivery of the message until a suitable message receive buffer becomes available.

An RDMA communication service may or may not provide a way for a sending consumer to determine whether a message has been delivered to the receiving consumer.

A.4 RDMA operations

An RDMA channel may provide RDMA Write and/or RDMA Read operations between its consumers. An RDMA operation allows one consumer associated with an RDMA channel, the requesting consumer, to transfer data to (RDMA Write) or from (RDMA Read) a memory address space belonging to the other consumer associated with the RDMA channel, the responding consumer.

Memory address spaces are identified by memory handles. Memory locations within a memory address space are identified by addresses. A responding consumer may obtain a memory handle by registering memory with an RDMA communication service. The responding consumer provides:

a. The range of addresses within the memory address space that will be valid.

b. Memory locations to be mapped to the valid addresses.

c. The RDMA channel or group of RDMA channels on which the resulting memory handle is to be valid.

d. Access control information, if required by the RDMA communication service.

e. Any other information required by the RDMA communication service.

to the RDMA communication service, which returns a memory handle if the registration is successful. The responding consumer then communicates the memory handle to the requesting consumer, which may use the handle to request RDMA operations.

An RDMA Write operation allows a requesting consumer to write data into a memory address space belonging to the responding consumer. A requesting consumer provides the following information to an RDMA communication service when it requests an RDMA Write operation:

a. The RDMA channel to use for the operation.

b. The memory handle for a memory address space.

c. A range of addresses within the memory address space.

d. Data to be written into the specified range of addresses.

An RDMA communication service may or may not provide a way for a requesting consumer to determine whether the data has been delivered to the responding consumer and written into the specified range of addresses. An RDMA communication service may or may not provide a way for a responding consumer to determine whether an RDMA Write operation is in progress or has completed.

An RDMA Read operation allows a requesting consumer to read data from a memory address space belonging to the responding consumer. A requesting consumer provides the following information to an RDMA communication service when it requests an RDMA Read operation:

a. The RDMA channel to use for the operation.

b. The memory handle for a memory address space.

c. A range of addresses within the memory address space.

d. A requestor buffer into which to place data read from the specified range of addresses.

The RDMA communication service notifies the requesting consumer after data has been successfully read from the specified range of addresses and placed in the requestor buffer. An RDMA communication service may or may not provide a way for a responding consumer to determine whether an RDMA Read operation is in progress or has completed.

A.5 Ordering and Reliability

SRP is designed to operate using an RDMA communication service having the characteristics described in this subclause. Operation with RDMA communication services having different characteristics may be possible, however any changes necessary to do so are outside the scope of this standard.

Communication on a single RDMA channel is ordered and reliable. The following properties are all true when an RDMA communication service notifies a receiving consumer that a message has been received on an RDMA channel:

a. Reliable messages: the received message has been received error-free and in its entirety.

b. Ordered messages: all messages that were sent to the same receiving consumer on the same RDMA channel and were sent prior to the message just received have already been received. The receiving consumer has already been notified of their reception.

c. Reliable and ordered RDMA Writes: all RDMA Write operations that referenced the receiving consumer as the responding consumer, used the same RDMA channel as the message just received and were requested before the message just received was sent have completed. The data for all such RDMA Write operations has been transferred completely and error-free, and is resident in the responder's memory address space(s).

An RDMA communication service may disconnect an RDMA channel if it is unable to satisfy these properties.